DETERMINISTICALLY ANNEALED MIXTURE OF EXPERTS MODELS FOR STATISTICAL REGRESSION

Ajit Rao[1], *David Miller*[2], *Kenneth Rose*[1], *Allen Gersho*[1]

[1]Dept. of Electrical and Computer Engineering,
University of California, Santa Barbara, CA 93106.
[2]Dept. of Electrical Engineering,
Penn. State University, University Park, PA 16802.

ABSTRACT

A new and effective design method is presented for statistical regression functions that belong to the class of mixture models. The class includes the hierarchical mixture of experts (HME) and the normalized radial basis functions (NRBF). Design algorithms based on the maximum likelihood (ML) approach, which emphasize a probabilistic description of the model, have attracted much interest in HME and NRBF models. However, their design objective is mismatched to the original squared-error regression cost and the algorithms are easily trapped by poor local minima on the cost surface. In this paper, we propose an extension of the deterministic annealing (DA) method for the design of mixture-based regression models. We construct a probabilistic framework, but unlike the ML method, we directly optimize the squared-error regression cost, while avoiding poor local minima. Experimental results show that the DA method outperforms standard design methods for both HME and NRBF regression models.

1. MIXTURE OF EXPERTS REGRESSION

In recent years, there has been growing interest in learning methods for regression functions that can be statistically interpreted as mixture models or mixture of experts (ME) models. The ME regression function takes the form:

$$g(\mathbf{x}) = \sum_j P[j|\mathbf{x}] f(\mathbf{x}, \Lambda_j), \qquad (1)$$

where $P[j|\mathbf{x}]$ is a non-negative weight of association between input, $\mathbf{x}$ and the jth "local expert regression function", $f(\mathbf{x}, \Lambda_j)$. Each local expert, $f(\mathbf{x}, \Lambda_j)$, is usually a constant, linear or simple nonlinear function of $\mathbf{x}$ and depends on the parameter set, Λ_j.

The weights of association can be naturally interpreted as a *probability distribution* since $\sum_j P[j|\mathbf{x}] = 1$. Hence, the ME model can be interpreted as a probabilistic partition of the input space - every point in the input space belongs in probability to partition cells, each of which is governed by a local regression model.

Some popular neural network approaches to regression, such as the hierarchical mixture of experts (HME) [4] and normalized radial basis functions (NRBF) [8], can be formulated as mixture of experts models.

1.1. The learning problem

Consider the problem of learning a regression function from a "training set", $\mathcal{T} \equiv \{(\mathbf{x}_i \mathbf{y}_i)\}, i = 1, 2..N$. The natural choice of learning criterion is the minimization of the average squared-error cost measured over the training set,

$$\min_{\{\Lambda_j\},\{P[j|\mathbf{x}_i]\}} D = \frac{1}{N} \sum_i \|\mathbf{y}_i - g(\mathbf{x}_i)\|^2 \qquad (2)$$

However in [3] and [4], a maximum likelihood (ML) criterion,

$$\max_{\{\Lambda_j\},\{P[j|\mathbf{x}_i]\}} L = \sum_i \log \sum_j P[j|\mathbf{x}_i] e^{-\|\mathbf{y}_i - f(\mathbf{x}_i, \Lambda_j)\|^2}, \qquad (3)$$

was preferred for a few reasons - Firstly ML training is fast, and performed better than minimum squared-error training in experiments. Further, ML training can realized by the Expectation Maximization (EM) algorithm [2] which has useful convergence properties. Also, ML training yields "competitive" solutions in which

This work was supported in part by the National Science Foundation under grant no. NCR-9314335, the University of California MICRO program, ACT Networks, Advanced Computer Communications, Stratacom, DSP Group, DSP Software Engineering, Fujitsu, General Electric Companuy, Hughes Electronics, Intel, Moseley Associates, National Semiconductor, Nokia Mobile Phones, Qualcomm, Rockwell International, and Texas Instruments. David Miller was supported by NSF Career Award NSF IRI-9624870

relatively few experts are strongly activated for any given input; Squared-error training yields "co-operative" solutions, where many experts typically contribute to give an output. In [3], competitive models were favored based on the advantages of a localized representation.

However, despite the advantages and promising results of the ML algorithm, we believe that the minimum squared-error cost is a more appropriate training criterion. Specifically, we note that the ML criterion of (3) encourages an *individual* fit between output $\mathbf{y}_i$ and each expert $f(\mathbf{x}_i, \Lambda_j)$, rather than the co-operative fit based on the ME output $g(\mathbf{x}_i)$. While gradient ascent on the ML cost surface sometimes minimizes the squared-error better than direct *gradient descent* on the squared-error cost, all that this really suggests is that the squared-error surface is more complex than the ML surface, with numerous poor local optima to trap simple descent methods. Rather than abandon the squared-error criterion to avoid the design difficulty, we suggest a more powerful method, deterministic annealing (DA), for the minimization.

2. DETERMINISTIC ANNEALING

Our work is based on the DA method proposed in the context of data clustering[11] and related problems[12] and its extensions to incorporate structurally constrained data clustering problems [5][6] which finds practical use in the design of statistical classifiers [7] and Generalized Vector Quantizers [9] for source-coding applications. The DA method is based on the interpretation of a mixture model as a radomized space partition. The degree of randomization of this partition can be measured by the Shannon entropy,

$$H = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j} P[j|\mathbf{x}_i]\log P[j|\mathbf{x}_i]. \qquad (4)$$

We first pose the problem of optimizing the regression cost, D of equation (2), while constraining the Shannon entropy, $H = H_0$. This constrained optimization problem may be written as the minimization of the corresponding Lagrangian,

$$\min_{\{P[j|\mathbf{x}]\},\{\Lambda_j\}} D - TH \qquad (5)$$

where the Lagrange parameter, T is referred to as the "temperature" to emphasize a compelling analogy to statistical physics. Equation (5) reminds us of the definition of the Helmholtz free energy of a thermodynamic system, where D is the thermodynamic energy of a physical system, T is the temperature and H is the entropy. The temperature (Lagrange multiplier) determines a balance of energy (cost) and entropy (randomness). Minimizing the Lagrangian, $D - TH$, we minimize the Helmholtz free energy, and in fact, seek isothermal equilibrium at the given temperature, T. Of particular importance is the case of $T \to 0$ which corresponds to direct minimization of D, our ultimate objective. This suggests the possibility of implementing an annealing process, that is, gradually lowering the temperature while maintaining the system at thermal equilibrium. Such a process allows one to avoid many of the local minima of the energy D. Since this method is not a stochastic method like simulated annealing, but instead based on the optimization of the deterministically computed expectation of the Helmholtz free energy, it is considerably faster.

We initialize the algorithm with a very high value of T. At this temperature, we must maximize the entropy of associating inputs with regions. The solution chooses all the probability distributions to be uniform and all the local models,$\{\Lambda_j\}$ to be equal to a single global model of the data. Effectively, a single region would suffice to represent the entire data. As the temperature is gradually lowered, in steps, optimization is carried out at each temperature to choose the parameters of the probability distribution and the local model parameters, $\{\Lambda_j\}$ that minimize the Lagrangian. As $T \to 0$, the Lagrangian reduces to the regression cost, D.

2.1. The NRBF structure

In the NRBF architecture,

$$P_j(\mathbf{x}) \propto e^{\frac{\|\mathbf{x}-\mathbf{m}_j\|^2}{2\sigma^2}} \qquad (6)$$

defines the association probability which depends on the relative closeness to each of the "prototypes", $\{\mathbf{m}_j\}$. Furthermore, $f(\mathbf{x}, \Lambda_j) = \Lambda_j$, i.e. we make use of constant local models. To design an NRBF regression function from the training data, one must optimize the locations of the prototypes, $\{\mathbf{m}_j\}$ and the local models, $\{\Lambda_j\}$ to minimize the regression cost, D. A common design approach[8] adopted for the design of NRBF-based regression functions can be summarized in two-steps :

- Find the prototypes, $\{\mathbf{m}_j\}$ that minimizes a clustering (VQ-like) cost in the $\mathbf{X}$ space.
- With this initialization for $\{\mathbf{m}_j\}$, use a gradient descent algorithm on the regression cost, D, of equation (2) to optimize $\{\mathbf{m}_j\}$,$\{\Lambda_j\}$.

While this algorithm is quick, a significant problem with it is that despite the heuristically reasonable initialization that the first step offers to the second (cost

minimization) step, the method can be easily trapped in poor local minima on the complex cost surface. We will demonstrate this problem with some examples in the next section and show that the DA approach effectively overcomes this shortcoming to generate better regression functions.

2.2. The HME structure

The hierarchical mixture of experts (HME) regression function is organized as a tree. The leaves of the tree represent simple local regression models (experts) which are weighted and combined, as they "traverse" to the root node, where the final regression estimate, $g(\mathbf{x})$ is computed. As an example, for a simple two-level, binary-tree HME architecture is [1], the output

$$g(\mathbf{x}) = \sum_j g_j(\mathbf{x}) \sum_k g_{k|j}(\mathbf{x}) f(\mathbf{x}, \Lambda_{jk}), \qquad (7)$$

where $g_j(\mathbf{x}) \propto e^{\mathbf{v}_j^T \mathbf{x}}$ and $g_{k|j}(\mathbf{x}) \propto e^{\mathbf{v}_{jk}^T \mathbf{x}}$ define the probability distributions. This architecture can be viewed as a mixture model, where each local model, $f(\mathbf{x}, \Lambda_{jk})$ is selected based on a probability that is the (tree-structured) product of the probabilities, $g_{k|j}(\mathbf{x})$ and $g_j(\mathbf{x})$.

Jordan and Jacobs[4] suggested the maximization of the likelihood function (3), as an effective method to design HME regression functions. However, although this method does sometimes minimize the regression cost better than gradient descent on the squared-error cost function, its design objective is mismatched to the original regression cost. In the results section, we will demonstrate that our novel DA method significantly outperforms both likelihood maximization, and gradient descent on the squared-error cost.

3. RESULTS

We applied our DA-based design method to the HME and NRBF architectures and compared the average squared-error performance with those obtained by conventional design methods for each architecture. In this section, we demonstrate that the DA method clearly outperforms the Two-step (2ST) method for the NRBF architecture and outperforms both the maximum likelihood (ML) and gradient descent (GD) methods for the HME architecture. The experiments are performed for different values of K, the number of local models used for regression. Note that in the case of NRBF regression functions, K is the number of prototypes and in the case of binary HME trees with l levels, $K = 2^l$. Since, for both HME and NRBF structures, the performance of the competing methods depends on the initialization used, we attempted to remove the bias introduced due to poor initialization by allowing each competing method to use ten different initializations with only the best result obtained among those runs compared with the result obtained by DA. Since the regression function obtained by DA is generally independent of the initialization, a single DA run will suffice. Further, in the case of HME regression functions, we use linear local models.

Our experiments are performed over three benchmark examples from the StatLib dataset archive. [2]

- **The Boston home value prediction problem** : The goal is to use a training set of data from 506 homes in the Boston area to predict the median price of each home from 13 features which are believed to influence it. Since the features have different dynamic ranges, we first normalize each feature to unit variance before designing regression functions for them. Results are shown in tables 1 (a) and (b).

- **Prediction of mortality rate** : We consider the prediction of the age-adjusted mortality rate in a locality from 15 factors that may have possibly influenced it. Since we have data on only 60 localities, we used the entire dataset for training. Results are shown in tables 2 (a) and (b).

- **Estimation of fat content of meat** : The fat content of meat can be measured by techniques of analytical chemistry, but it is a slow and time-consuming process. In this experiment, we used a dataset of quick absorption measurements [3] and the corresponding fat content as determined by analytical chemistry to learn to predict the fat content from the measurements. The data consists of a training set of size 173 and a test set of size 43. The results using NRBF and HME regression functions for the prediction are shown in tables 3 (a) and (b).

Clearly, for all three examples, the deterministic annealing method outperforms the standard design methods for both mixture model architectures. Note that, in table 3(b), allowing the ML approach to use a larger network size does not necessarily improve the performance over the test set, although performance over the training set improves marginally.

[1] Note that our method is not restricted to binary or two-level trees. This example is given only for ease of understanding.

[2] The StatLib data set archive is accessible on the World-Wide Web at http://lib.stat.cmu.edu/data sets/ .

[3] The Tecator Infratec Food and Feed Analyzer measures the absorption of electro-magnetic waves in 100 different frequency bands

4. ACKNOWLEDGMENTS

The authors wish to thank Professors Michael I. Jordan and Prof. Robert A. Jacobs for providing their HME design software.

5. REFERENCES

[1] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., Classification and regression trees, Belmont, CA: Wadsworth, 1984.

[2] Dempster,A.P., Laird,N.M., Rubin,D.B., Maximum-likelihood from incomplete data via the EM algorithm, Journal of the Roy. Stat. Soc., Ser. B, 1977, V 39, p1-38.

[3] Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E., Adaptive mixtures of local experts, Neural Computation, Spring 1991, vol.3, (no.1):79-87.

[4] Jordan, M.I., Jacobs, R.A., Hierarchical mixtures of experts and the EM algorithm, Neural Computation, March 1994, vol.6, (no.2):181-214.

[5] Miller, D., Ph.D. thesis, University of California, Santa Barbara, 1995.

[6] Miller, D., Rao, A., Rose, K., Gersho, A., An Information-theoretic Learning Algorithm for Neural Network Classification, Neural Information Processing Systems, pp. 591-597, 1995, vol. 8.

[7] Miller, D., Rao, A., Rose, K., Gersho, A., A Global Optimization Technique for Statistical Classifier Design, IEEE Transactions on Signal Processing, December 1996.

[8] Moody, J., Darken, C.J., Fast learning in networks of locally-tuned processing units, Neural Computation, Summer 1989, vol.1, (no.2):281-94.

[9] Rao, A., Miller, D., Rose, K., Gersho, A., A generalized VQ method for combined Compression and Estimation, Proceedings ICASSP-96, pp. 2032-2035., vol.4.

[10] Rao, A., Miller, D., Rose, K., Gersho, A., An Annealing Approach to Parsimonious Modeling in Statistical Regression, submitted for publication.

[11] Rose, K., Gurewitz, E., Fox, G.C., Vector quantization by deterministic annealing, IEEE Transactions on Information theory 38, 1249-1258, 1992.

[12] Rose, K., Gurewitz, E., Fox, G.C., "Constrained clustering as an optimization method", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, pp. 785-794, 1993.

K	DA	2-Step
1	87.7	87.7
2	19.7	23.78
4	12.9	19.34
6	12.6	13.74
10	6.5	15.72

K	DA	GD	ML
4	5.73	5.88	7.49
8	3.40	3.58	5.59

Table 1: Comparison of average squared-error for the Boston home data problem using (a) NRBF architecture and (b) the binary HME tree. K is the corresponding network size.

K	DA	2-Step
1	3805.12	3805.12
2	1148.82	2154.0
4	720.77	1256.8
6	439.07	566.5
8	299.63	564.5
10	261.42	438.2

K	DA	GD	ML
4	18.2	121.8	70.4
8	2.1	12.3	41.8

Table 2: Comparison of average squared-error for the environmental data problem using the (a) NRBF architecture and (b) the binary-HME tree. K is the corresponding network size.

K	DA (tr)	DA (te)	2-Step (tr)	2-Step (te)
1	159.89	168.25	159.89	168.25
2	52.9	58.8	131.43	159.68
4	28.6	32.9	119.82	137.99
6	27.3	40.1	74.89	83.73

K	DA (tr)	DA (te)	GD (tr)	GD (te)	ML (tr)	ML (te)
4	8.3	11.5	14.1	18.1	15.1	23.9
8	6.9	9.8	12.8	17.2	12.5	39.7

Table 3: Comparison of average squared-error for fat-content estimation using (a) the NRBF architecture and (b) the binary HME tree. K is the number of basis functions. 'tr' and 'te' refer to training and test sets respectively.